Introduction

Town centers form the core of many urban areas and are characterized by clustering of various types of socio-economic activities with retail and related services being pivotal. They can be viewed as complex economic systems that constantly evolve (Thurstain-Goodwin and Unwin 2000) and therefore their composition and spatial extent are likely to expand or contract over time. This evolution has been linked to changes in the planning system, rising property values, changing levels of accessibility, other forces of change such as economic shocks or more gradual changes such as the rise of Internet sales (Singleton et. al. 2016).

It has long been recognized within multiple international settings, that the aggregate national structure of consumer spaces and shopping destinations are complex (Berry 1967); with retail cluster size and function relating to their attraction, market potential, competition and agglomeration benefits. Within many contexts traditional shopping destinations that have evolved naturally and appear well-embedded within the urban fabric (including town centers), are supplemented by purpose-created retail opportunities such as regional shopping centers, retail parks, strip malls or focused shopping destinations such as designer outlets (Teller and Reutterer 2008). Although it has been argued that depicting retail agglomerations for a national extent, and particularly accounting for more granular temporal shopping patterns is very challenging (Mackaness and Chaudhry 2011); the classification of shopping destinations and delineation of their spatial extent is essential to gaining a better understanding of the relationship between use of retail space and changing consumer behavior (Guy 1998). A consistent and rigorous approach to defining town center boundaries enables systematic metrics of retail center morphology and performance to be actualized (Thurstain-Goodwin and Unwin 2000), alongside providing utility as input into many commonly implemented retail analytics tasks related to store location and demand estimation (Newing et al. 2015).

In the case of England and Wales, a national set of town center boundaries were developed by Thurstain-Goodwin and Unwin (2000) and subsequently adopted by the Department of Communities and Local Government (DCLG) in 2004. Their approach was to generate surfaces of spatial densities using kernel density estimation, from socio-economic variables including building density, diversity of building use, and tourist attraction (Mackaness and Chaudhry 2011). In addition, their approach aimed at delineating town centers, however, such zones are more expansive (e.g. by including office space) than those that might be related mainly to retail. As such, one of the objectives of this work is to move away from a more general definition of town center locations as centers for employment, to a more functional measure of spaces delineated for retail and services. Furthermore, in many cases, the extent of the 2004 DCLG town center boundaries will likely have changed over the past decade, eroding the utility of these previous models for contemporary applications. Finally, the availability of more accurate and comprehensive spatial data on retail unit locations in Great Britain (G.B.) has improved significantly since this time, which provides scope for exploring a new robust method of defining the spatial extent of retail agglomerations. As such, this paper highlights deficiencies in a number of existing cluster analysis methods for retail center definition before presenting a density-based clustering technique that can consistently identify retail areas, is updatable over time and can be applied to wider national extents. We implement this analysis using a national dataset of retail and service locations, and evaluate the center definition outcomes at a local level.

Where are retailers located?

A national occupancy dataset of 529,062 retail locations across G.B. was provided by the Local Data Company (LDC) through the ESRC Consumer Data Research Centre and was collected via a large pool of local surveying teams during 2015. The data contain detailed information about the current occupier and location of retail unit and service premises.

While a full postcode was available for all surveyed premises (enabling geocoding proximal to ~13 properties), more precise latitude and longitude coordinates were available for 437,260 units (about 82%), which were retained for further analysis; thus providing building level of accuracy. Other collected information for each location included the fascia (a surrogate for occupier) and the type of retail or service business (i.e. leisure, comparison, service and convenience) including vacant outlets. For retail units located in shopping centers, retail and leisure parks the respective name of the shopping center or retail park was also provided.

Conceptually, utilizing vacant units in the identification of local retail agglomerations may be problematic given that these voids may often occur as a result of failure of a particular retail setting (Benjamin et al. 2000), and as such, an indication of potential change in extent morphology. For this reason, all vacant units were removed from the dataset. Additional processing also removed units that were classified as auto services that are not typically considered part of retail agglomerations. Furthermore, miscellaneous (not related to retail or unclassified units) were also excluded. The final cleaning operation identified and removed duplicate locations (i.e. points with identical coordinates or within very close proximity), which can unduly influence clustering results as well as the identification of outliers. These duplicate locations were typically the result of the two-dimensional representation of retail units within multi-storey buildings. Thus, the removal of duplicates (any points within a 2 meter radius from another point) was carried out.

Estimating retail center location and extent; methods and calibration

Cluster analysis is a collection of unsupervised learning methods that address the issue of grouping a set of objects based on similarity. Many commonly used clustering algorithms make group allocations with the objective of increasing similarity within a cluster and increasing dissimilarity between clusters. Other commonly used clustering techniques such as density-based algorithms seek dense regions separated by low density regions,

while model-based methods assume that the data come from a mixture of probability distributions, each of which represents a different cluster (Gan et al. 2007). Cluster analysis is a multivariate technique (multiple attributes of the phenomenon under investigation can be used), but in this study it is strictly spatial; utilizing only the locations of the retail units. This is an appropriate approach for the identification of retail agglomerations where the extent of the clusters are determined by spatial discontinuity in unit distribution (Dearden and Wilson 2011).

An important consideration when clustering spatial data is to select a method that is sensitive to the distinction between clusters that are either compact or chained (Gan et al., 2007) and, additionally, can identify outliers outside of primary observed geographic distributions. Within a retail context, examples of compact clusters could include those retail units residing within a city or major town center such as Wolverhampton (West Midlands) (Figure 1A), often with connecting voids that are pedestrianized. Chained retail clusters on the other hand often can be observed along the road network (these are often known as “high streets” in Great Britain), such as Clapham Junction (London) (Figure 1B).

[Figure 1A and Figure 1B here]

In order to estimate the definition of retail centres, the following clustering methods were evaluated: DBSCAN (Ester et al. 1996), Quality Threshold (Scharl and Leisch 2006), Kernel Density Estimation (Azzalini and Torelli 2007), Random Walk (Csardi and Nepusz 2006) and K-means (Lloyd 1982). As will be described, all of the clustering methods evaluated require the calibration of tuning parameters that we selected to optimize using the S_Dbw internal evaluation indicator (Halkidi and Varzigiannis 2002), which has been found by Liu et al. (2010) to provide better results compared to seven other internal validation indexes. It is defined as the sum of the mean dispersion (S) in the clusters and of the between-cluster density (G) (Desgraupes 2013):

$$S\_Dbw = S + G \quad (1)$$

As such, the process of calibrating each clustering method was carried out prior to implementation in the evaluation by identifying suitable starting values (for those tuning parameters that a single value could not be determined), then producing a number of different models within a range of values and finally selecting the optimal model based on the S_Dbw index (i.e. selecting the parameter values of the model with the smallest S_Dbw value).

DBSCAN (Density Based Spatial Clustering of Applications with Noise) (Ester et al. 1996) was selected as it is one of the most prevalently implemented spatial clustering algorithms that is able to find arbitrarily shaped clusters and to handle outliers (Gan et al. 2007). In addition, with the use of kd-tree indexing this was the computationally fastest method tested. The greatest drawback of DBSCAN is limited sensitivity for datasets with varying densities (Everitt et al. 2011). Our optimization for the epsilon (radius) parameter started by calculating the distance to the 4 nearest neighbors for each point (Ester et al. 1996). The distances were then sorted in ascending order and the 95^{th} percentile value was selected as starting epsilon. Even though this is a simple technique, k-NN distance has been found to be a reliable proxy of local density and outlier detection, outperforming even newer and more complicated methods (Campos et al. 2016). The minimum points parameter was set equal to 10, which is the minimum number of retail units required for an area to be classified as local center (Wrigley and Lambiri 2015). Following this, the DBSCAN method was calibrated by allowing the epsilon value to vary within the range of +- 20 meters from the starting epsilon value. Using 5 meter intervals, the best clustering solution from 9 DBSCAN models for every study area was selected with the S_Dbw index. Within the study sites, the 20 meters range was used as it was found to be large enough to test as many models as possible without being an extremely demanding task, while the 5 meters interval was small enough that any difference between models using a smaller interval was negligible.

N*on-parametric density estimation* (Azzalini and Torelli 2007; Azzalini and Menardi 2014) combines both kernel density estimation (KDE) and a graph model that connects retailers into a network by proximity. In this process, KDE is used to identify a number of core clusters with density above a certain threshold from within the spatial distribution. These are then used to create connected regions of points (subgraphs) by means of Delaunay triangulation. The technique requires definition of a parameter value that is multiplied by the smoothing vector of the kernel estimator. This was determined through comparison of retail boundaries delineated by respective local authorities to outputs created with the clustering method. Suitable values for the smoothing parameter varied between 0.4 and 1.1. Lower values resulted in too fragmented clusters, while higher values over-smoothed, creating large and also unrealistic clusters. Multiple models were tested, varying the smoothing parameter value using 0.05 intervals, and again selecting the optimal clustering with the S_Dbw index. A key advantage of this method is that it is non-parametric (it does not make any assumptions concerning the probability distribution), and thus is more suitable to identify clusters of varying shapes and densities. However, it is also a stochastic method, and as such it requires optimization, which has the disadvantage of increasing computation times.

The *Quality Threshold* (QT) (Scharl and Leisch 2006) identifies clusters after specification of two parameters: the maximum diameter of the clusters and the minimum number of neighbors within a cluster. The minimum number of neighbors was set equal to 10 which aligns to a formal definition of a retail center within the UK (Wrigley and Lambiri, 2015). Through testing within different contexts, the optimal radius value was highly sensitive to retail unit density variation. After consideration of the S_Dbw index, the radius parameter was allowed to vary between 100 and 400 meters with 50-meter intervals for smaller urban areas (e.g. Abertillery) and between 300 and 1000 meters with 100-meter intervals for larger urban areas (e.g. Bristol). The algorithm initializes by randomly selecting a point as

a center of a cluster and then, for as long as the diameter is smaller than a user specified value, it iteratively adds a point to the cluster so as to minimize the increase in the cluster diameter. This process is repeated for a random number of sample center points that satisfy the condition of having at least one neighbor within the specified diameter threshold. After the largest candidate cluster is identified and removed from the dataset the process is repeated for as long as there are no remaining clusters with size greater than the neighbor threshold. The method is also computationally intensive due to being stochastic.

Random Walk was tested which is a graph-based method that is based on the Walktrap algorithm (Pons and Latapy 2005). The algorithm finds densely connected subgraphs based on the assumption that random short walks tend to stay within the same densely connected subgraph. Initially, the algorithm partitions the graph into a number of subgraphs and then computes the distances to all adjacent vertices. Subsequently, for each iteration it chooses two subgraphs to merge if they are adjacent and if they minimize the squared distances between the vertices. The output is a dendrogram where the leaves are the vertices and each edge is a connection between subgraphs. The best partition of the graph is the one that maximizes a modularity criterion (Newman 2004). Optimization found that the method required a maximum number of 50 steps in order to find the best model using the S_Dbw index.

The final algorithm tested was *K-means,* with the only parameter requiring specification being the number of K clusters. Initially, the algorithm allocates objects randomly to each cluster and, subsequently, iteratively assigns the objects to the nearest cluster according to a distance measure until either the distance measure or the membership of the clusters do not change significantly. This method has low computational complexity, however, produces clusters with convex hull shapes and it does not always identify outliers, that is, all objects are clustered although may return outlier clusters with very small case

frequency. In addition, the method is also stochastic, and therefore requires optimization through multiple runs which occurs at the expense of computational time. Information obtained from the application of the other clustering methods was used to calculate the starting value of the number of clusters as the mean number of the clusters identified by DBSCAN, KDE, QT and Random Walk. Subsequently, the method was calibrated by producing 11 models with the number of clusters varying within the range of +- 5 clusters from the starting value and the optimal model was selected based on the S_Dbw index.

In additional to the aforementioned methods, the Chameleon (Karypis et al 1999), Fast Greedy (Clauset et al. 2004) and Ensemble (Hornik, 2007) methods were also tested but are not used for the evaluation. Chameleon was not included given difficulty in automating the process of identifying optimal values for its six tuning parameters, Fast Greedy is a graph-based method that did not provide better results than the Random Walk and finally the Ensemble method was particularly demanding in terms of computer resources for a nationally extensive application. Obviously, there are a plethora of other methods that have been shown to be useful for clustering spatial data such as the DBCLASD method (Xu et al. 1999). However, an important factor for inclusion in the evaluation was that the methods were accompanied by useful documentation that facilitated their implementation. In addition, that there was indication they were are under active development or well established, and were available within most programming languages.

Center definition and evaluation

The five candidate methods were evaluated over eight case study areas that are representative in terms of G.B. retail location density and size. These included: Abertillery and Cardiff in Wales, Bristol, Clapham Junction, Winchester and Wolverhampton in England, Glasgow and Inverurie in Scotland (Figure 2 and Figure 3).

[Figure 2 and Figure 3 here]

Although there is a larger pool of other representative areas, within these specific locations additional supplementary data were also available for cross validation and included two sources. Firstly, local authorities within the U.K. are required to perform a town center “health check” (NPPF 2012), which typically requires them to delineate boundaries for retail centers. Even though the reports produced by the local authorities contain rich information, the publicly available boundaries can typically only be accessed in rendered pdf format. Given that a small number of (qualitative) comparisons can be made against these sources without extensive re-digitizing, the reports were used to assist with input parameter specification and testing during the calibration process described in the previous section. Secondly, boundaries for the 339 largest “retail places” in the U.K. were acquired from the company Geolytix, and although they represent only a subset of total retail boundaries, they nevertheless provide an additional and relatively large sample of independent retail areas suitable for comparison.

Finally, within evaluation that follows, all clusters (identified by each clustering method) that had less than 10 retail units were removed, which as noted earlier, is the minimum threshold considered to be as part of a center. Additionally, for those clustering solutions that additionally identified outliers from the main distributions, these locations were also removed.

The remainder of this section presents the outputs of the clustering methods for two of the larger more complex study areas: Bristol and Glasgow, alongside an overall set of evaluation results for all case study locations.

Bristol has a greater than average number of retail units (2456), high variability of retail density and potential occurrence of different cluster shapes. The location of the retail units (blue dots) are shown in Figure 4 alongside labels colloquially used for the various retail centers and their boundaries as defined by the respective local authorities.

[Figure 4 here]

In the past, Broadmead was recognized as the principal shopping center of Bristol, but recent studies (Bristol City Council 2008 [unpublished]) suggest that the boundaries of Bristol should be expanded to include the high streets of Stokes Croft south of Ashley Road (depicted as sparse dots in Figure 4), Christmas Steps and Old Market. The most recent Local Plan from 2015 (BCAP, 2015), which is required by law, defines precisely the boundary of a wider Bristol city center, however the spatial extent of the individual so-called shopping, services and the evening economy areas is less specific as these areas often have overlapping functions. The Local Plan defines the primary shopping area as Broadmead and Queen's Road; in addition, it defines the primary shopping frontages (Broadmead and part of Queen's Road and Old City), secondary shopping frontages Stokes Croft, Old Market, Victoria Road and parts of Queen's Road and Old City and leisure use frontages (part of Old City and Broadmead).

The first clustering algorithm to be evaluated was DBSCAN, which identified 26 clusters in the study area as can be seen in Figure 5 (outliers are denoted by 0). Stokes Croft is part of the city of Bristol (with the cluster boundary extending north of Ashley road), however, Old Market is not. There is a good separation from the Gloucester Road cluster that has been identified correctly as a single cluster. Clifton, Whiteladies and most of the town center have also been identified as separate clusters. Within Bedminster, the western part of the area was however identified as a separate cluster, most likely due to higher local density. The KDE method identified 11 clusters with a cluster for the city center being fairly accurate, matching the local authority defined boundary. However, it is obvious that the method identified fewer clusters than might be expected given the overall retailer distribution. The clustering solutions generated by QT, K-means and Random Walk were somewhat similar in that they identified separate clusters in areas that are strongly connected (e.g. Bristol city center, Gloucester road) while they clustered together points that are weakly connected (e.g. Totterdown and Well road for QT and K-means, Queen's

road and Clifton for Random Walk). A further problem with the methods is that they identified few outliers, which results in the identification of very sparse clusters.

[Figure 5 here]

With 2347 retail units, Glasgow it is the second largest study area in the analysis (Figure 6). There is one metropolitan retail center (Glasgow city), one regional center (Partick – Byres road) and 5 town centers (Calton, Crastonhill – Yorkhill, Kelvinbridge, St. George's Cross – Great Western road and Woodlands) (Figure 6). The boundary of Glasgow city is well defined by the M8 motorway (north and west), the river Clyde (south) and the High street (west).

[Figure 6 here]

DBSCAN (Figure 7) identified accurately the cluster of Glasgow city center, with only a few retail units crossing the M8 on the west of the city and south of Woodlands. The QT and KDE methods clustered the city center together with the town center of Calton. K-means and QT also merged the western part of Glasgow city with Woodlands and St. George’s Cross. The output from the Random Walk had additional issues, splitting up the larger retail areas as in the case of Partick-Byres road. For that retail area, DBSCAN provided the most accurate result, however, the boundary of the cluster extended to include Kelvinbridge. Concerning St. George's Cross, the cluster obtained from DBSCAN is a close match to the boundary defined by the Glasgow city council and the same could be said for Woodlands.

[Figure 7 here]

Table 1 presents the overall evaluation results from the qualitative comparison for all of the eight study areas. In most cases, the DBSCAN method provided results that were more consistent with those formal definitions created from the respective local authorities. Importantly, DBSCAN was the most efficient method in terms of computing resources and this is particularly significant for a national extent study. In addition, it was easier to identify

starting values for the parameters of the method, while one of the strongest advantages of DBSCAN was the identification of outliers.

[Table 1 here]

It is clear from the results that DBSCAN performed well for the case study selection, however, this method is known to underperform in areas where the density is not uniform (Everitt et al. 2011). Such an issue also becomes apparent when looking at the range of the optimal epsilon values that were used for the selected areas (Table 2). If a single global epsilon value had been used for all case studies, it would have resulted in suboptimal local results. As such, we developed a refinement to the method which involves splitting of the national-scale data into more homogeneous areas for separate treatment; with the challenge being that unlike the case study evaluations, this required automation given that coverage was for the national extent.

[Table 2 here]

Development and application of a modified DBSCAN method

In order to address the issue of heterogeneous density, a modified approach to DBSCAN was developed by introducing three important concepts:

(1) the combination of DBSCAN with graph data structures and algorithms that are used to iteratively partition the national study area into subgraphs of successively more homogeneous point density;

(2) the iterative application of DBSCAN using a local epsilon value for each subgraph, followed by the selection of one cluster per iteration based on the condition that the epsilon value is representative of the cluster’s density;

(3) the use of a third parameter termed maximum distance to constrain the points that can be members of a cluster to have at least one neighbor within a radius that is less than or equal to the maximum distance. The rationale behind this decision is that distance is an important parameter of retail spatial agglomerations, which is sensitive to gaps and

discontinuities. Given that both spatial density and spatial discontinuity determine whether a point is part of a spatial cluster, the combination of k-nearest neighbors (a proxy of point density) with the radius-based constraint (a proxy of spatial discontinuity) facilitates neighboring locations within close proximity and similar point density to be members of the same cluster. Compared to a post-processing removal of points based on a distance threshold, using a distance threshold within the modeling process has the advantage of avoiding the inclusion of outliers in the calculation of the epsilon value but, more importantly, facilitates the decomposition of a graph into subgraphs of more homogeneous density.

In the first step of the proposed methodology, a sparse graph representation of the spatial dataset is created based on a k-nearest neighbor matrix and the maximum distance constraint. The vertices of the graph are the locations that have at least one neighbor within the specified maximum distance. Next, a Depth First Search algorithm is implemented to decompose the sparse graph to create more homogeneous (in terms of point density and distance between the retail units) subgraphs, under the condition that each subgraph has at least 10 vertices and that each location has at least one neighbor within the maximum distance. The vertices that are not part of any subgraph are removed as outliers. The maximum distance value in this study represents the maximum distance that a location can still be considered well connected to a shopping area on foot. Different distance values have been suggested as indicators of walking distance, ranging between 300 to 500 meters (NPPF 2012; Rogstad and Dysterud 1996). Based upon the definition of edge of center for retail purposes in the UK (DCLG, 2009), the maximum distance value was set equal to 300 meters. Three k values were tested to split the study area into subgraphs, and included 4, 10 and 15 (Figure 8). The first value was tested as it is already used as a proxy of local density and the second value was considered as it is used by the minimum points parameter of DBSCAN. As it would be expected, the lower the k value the

greater the number and the more homogeneous the density of the subgraphs that were produced. On the other hand, using lower k values (between 4 and 10) can result in splitting areas with low point density (mostly chained clusters, i.e. High Streets) into different subgraphs. For this reason the k value was set equal to 15.

[Figure 8 here]

Given that the spatial extent of each subgraph depends on the connectivity and number of points within an area, each subgraph can represent a town center, a city center or even a metropolitan region. DBSCAN, however, assumes that the epsilon value is a representative indicator of the local density. To fulfill that assumption, in the third step of the methodology, DBSCAN is first applied (within each subgraph) in an exploratory approach to identify and select the cluster that has density (as estimated by the local epsilon, i.e. the 95^{th} percentile of the 4-nearest neighbors' distances) closer to the overall density.

Following the selection of a single cluster, all the neighboring clusters (i.e. the clusters that share a common edge in the graph) with similar density are selected along with those neighboring points that were identified by the exploratory DBSCAN as outliers. Following this, a new study area of homogeneous point density is created from the selected points and DBSCAN is applied again to identify the clusters. The selected clusters are then removed from the graph representation of the point data, and the process of using an exploratory DBSCAN model to identify a cluster and select those neighboring clusters with similar point density is iteratively carried out until no cluster can be formed. This process is summarized in Figure 9. It should be noted that one of the advantages of the methodology is that it is no longer required to optimize the clustering solution using the S_Dbw index, which results in a faster algorithm.

[Figure 9 here]

To evaluate the point density similarity among clusters, the standard deviation of point density in a subgraph was used. More specifically, those neighboring clusters with point density within 1 standard deviation from the point density of the initially selected cluster were also selected, with the assumption being that they define an area of homogeneous point density. To test the sensitivity of the method to the standard deviation threshold, five different values were considered, 0.6, 0.8, 1.0, 1.2 and 1.4. As can be seen in Table 3, the clustering solutions are practically identical when looking at the number of clusters produced and the distribution of the local epsilon value.

[Table 3 here]

For the parameter values required by DBSCAN, as detailed earlier, the value of the minimum points parameter was set equal to 10 and the epsilon value was calculated as the 95^{th} percentile of the 4-nearest neighbor distance. However, the epsilon value was only allowed to vary within the range between maximum 170 meters, which was found to be useful to exclude outliers from being identified as members of clusters, and a lower bounds of 80 meters which was used to avoid identifying certain large shopping malls as clusters. This necessity is a consequence of the hierarchical nature of retail centers within G.B. given that the objective of the analysis was to create clusters that were inclusive of the different functional retail forms. Following the application of DBSCAN to each subgraph and the extraction of 2920 clusters, the final retail agglomerations were compiled and each retail location was assigned an identifying number denoting cluster membership. The clusters obtained from the modified DBSCAN methodology for the selected study areas are shown in Figure 10 and can be compared against those created by applying the traditional DBSCAN to each subgraph (Figure 11). For the traditional DBSCAN model a global epsilon equal to 107 meters was applied, which was calculated as the 95% of the 4-nearest neighbors distance.

[Figure 10 and Figure 11 here]

When comparing the two graphs, it can be seen that in certain areas such as Bristol and Cardiff the clustering solutions are quite similar, however, in areas such as Clapham Junction and Wolverhampton the modified DBSCAN model appears to be more sensitive to gaps and discontinuities, thus identifying a greater number of clusters. Particularly for Glasgow, the modified DBSCAN method provided the only clustering solution that identified Kelvinbridge as a separate cluster in an area of high point density that does not provide major discontinuities between clusters. At the same time, it was the only method that identified a sparse cluster south of the river Clyde and west of the M8 motorway (the epsilon value was 80 meters for Kelvinbridge and 170 meters for the cluster south of Glasgow). Similarly, for Inverurie, the modified DBSCAN method used an epsilon value of 170 meters to correctly identify a single cluster in the study area, compared to the two clusters identified by the traditional DBSCAN method when the global epsilon value of 107 meters was used.

The results derived with this new method were compared to data supplied by the company Geolytix; which represent the only freely available and independently created national sample of contemporary retail center extents. They provide frequent updates of a dataset of retail places across the U.K., part of which (339 places) were licensed as open data in 2012. The Geolytix boundaries are produced using multiple variables (including the locations of retail units) (OpenData 2015) with information that was collected at least three years prior to the data that were used in our analysis. Additional causes of difference between the two datasets might also include the different objectives and notion of what constitutes a retail center (Geolytix did not use a threshold of minimum 10 retail units), and only the boundary polygons from the clustered locations of the retail units were available. Given that the creation of similar polygon boundaries for our new results may result in an additional source of error, it was decided to compare the Geolytix boundaries against the retail unit locations and associated clusters. The comparison was based on two metrics,

the n-ary relation between the two datasets and the proportion of points within the Geolytix polygons. The n-ary relation returns a score where the higher the number of clusters that had one-to-one relation with the clusters identified by Geolytix the better the relation.

Data pre-processing removed the major out of town retail parks from the Geolytix dataset, which was followed by a spatial join of the Geolytix dataset with the clustered retailer locations. There were 294 spatial intersections between the two datasets, out of which 244 were one-to-one. Summary values of the spatial distribution of the clustered locations within the Geolytix boundaries are shown in Table 3. On average (based on the median value) almost 90% of the clustered points were within the Geolytix boundaries.

Glasgow (Figure 12) serves as an example where the two datasets mostly overlap, but also shows that the spatial extent of the clusters produced in this analysis was on average larger, which to some extent is related to Geolytix post-processing of boundaries to be constrained by the road network. Examples where the two datasets have significant differences include Bristol (Figure 13) and London (Figure 14).

[Figure 12 and Figure 13 here]

Concerning Bristol, it can be seen that Geolytix split the city center into smaller clusters, of which only Broadmead was available as open data. However, the clustering solution for Bristol that was produced in this analysis was very similar to the one produced by the Bristol local authority and, thus, arguably more appropriate based on this local knowledge. Geolytix also split London into smaller clusters, 7 of which were available for the area that was identified by the modified DBSCAN method as a single cluster. A possible reason for this difference could be that Geolytix used additional variables in their clustering method, which, particularly for London, would result in identifying clusters based on different retail activities rather than just retail density. Despite these mismatches that to some extent are related to different objectives and notions of what constitutes a retail center, it could be argued that the two clustering solutions largely overlap in the areas that were available by

the open source Geolytix retail places, which provides evidence for the validity of the retail clusters that were produced in this work vis-à-vis competing methods.

[Table 4 here]

Conclusion

The objective of this analysis was to develop a clustering method that would facilitate the identification of retail agglomerations across a national extent and that could be updated over time. For this purpose, five of the most frequently used clustering methods were compared within 8 representative locations across Great Britain. The DBSCAN method was selected on the basis that it provided the most accurate representation of those retail areas relative to formal definitions; it was faster to produce a clustering solution and also easier to calibrate optimized input parameter values.

However, in order to address a well-known issue that DBSCAN does not cope well in areas of varying densities, the DBSCAN method was adapted so that it could be iteratively applied within smaller more homogeneous sites that were created using a k-NN sparse graph representation of the retail locations. Each selected retail cluster was created by the DBSCAN algorithm with an epsilon value that was representative of the local point density. The clusters produced were comparable to those retail areas designated by the local authorities for the sample areas of study, and in some cases, were more accurate when compared to the traditional DBSCAN method. In addition, the identified clusters were in most areas similar in terms of spatial extent to those produced by the Geolytix company using alternative dataset and methodology. It should be noted that even though the suggested method is more demanding in terms of computer resources compared to the traditional DBSCAN, it scales better as it could be applied in parallel for each subgraph.

Furthermore, the output of this analysis provides a better spatial coverage and option for automated update in comparison to the existing DCLG town center boundaries. Given that the DCLG boundaries were widely used by academics, local authorities and private

organizations across the country it can be anticipated that these results will prove to be valuable for research and analysis.

With the developed methodology being open source (github / data links will be added post review), it will also be straightforward to update the retail boundaries on a regular basis, and potentially apply the suggested method within a context of historic data. Finally, given the variety in point density, size and shape of the retail clusters in the dataset it would be reasonable to assume that the methodology could be applicable with different datasets and for different international locations.

References

Azzalini, A. and N. Torelli. (2007). “Clustering via nonparametric density estimation.” Statistics and Computing. 17, 71-80.

Azzalini, A. and G. Menardi. (2014). “Clustering via nonparametric density estimation: the R package pdfCluster.” Journal of Statistical Software, 57(11), 1-26.

BCAP. (2015). Bristol central area plan. Available at: https://www.bristol.gov.uk/documents/20182/34540/BCAP%20Adopted%20March%202015%20-%20Policies%20Map%20-%20Web%20PDF.pdf/49f1d2b3-dda4-4ecf-8cf9-e20fe7b7c34c

Benjamin, J.D., G.D. Jud and D.T. Winkler. (2000). “Retail Vacancy Rates: The Influence of National and Local Economic Conditions. Journal of Real Estate Portfolio Management.” 6(3), 249-258.

Berry, B.J.L. (1967). “Geography of market centers and retail distribution.” Englewood Cliffs, NJ: Prentice-Hall.

Campos, G.O., A. Zimek, J. Sander, R.J.G.B Campello, B. Micenkova, E. Schubert, I.

Clauset, A., M.E.J. Newman and C. Moore. (2004). “Finding community structure in very large networks.” Available at: https://arxiv.org/pdf/cond-mat/0408187.pdf

Csardi, G and T. Nepusz. (2006). “The igraph software package for complex network research.” InterJournal, Complex Systems, 1695. Available at: http://www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf

DCLG. (2009). “Practice guidance on need, impact and the sequential approach.” Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/7781/townce ntresguide.pdf

Dearden, J. and A. Wilson. (2011). "A Framework for Exploring Urban Retail Discontinuities". Geographical Analysis, 43(2), 172-187.

Desgraupes, B. (2013). “Clustering Indices.” Available at: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Ester, M., H.P. Kriegel, J. Sander and X. Xiaowei. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise.” Available at: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Everitt, B.S., S. Landau, M. Leese, and D. Stahl. (2011). “Cluster Analysis.” 5th ed. Chichester, U.K.Wiley.

Gan, G., C. Ma and J. Wu. (2007). “Data clustering. Theory, algorithms and applications.” ASA-SIAM Series on Statistics and Applied Probability.

Halkidi, M. and M. Vazirgiannis. (2002). “Clustering validity assessment using multi-representatives”. Available at: http://lpis.csd.auth.gr/setn02/poster_papers/237.pdf

Hornik, K. (2007). “A CLUE for CLUster Ensembles.” Available at: https://cran.r-project.org/web/packages/clue/vignettes/clue.pdf

Karypis, G., E. Han and V. Kumar. (1999). “Chameleon: Hierarchical clustering using dynamic modelling.” IEEE Computer 32(8): 68-75.

Liu, Y., Z. Li, H. Xiong, X. Gao and J. Wu. (2010). “Understanding of Internal Clustering Validation Measures”. Available at: http://datamining.rutgers.edu/publication/internalmeasures.pdf

Lloyd, S.P. (1982). “Least squares quantization in PCM.” IEEE Transactions on Information Theory 28, 128-137.

Mackaness, W.A. and O.Z. Chaudhry. (2011). “Automatic Classification of Retail Spaces from a Large Scale Topographc Database.” Transaction in GIS 15(3), 291-307

Newman M. E. J. (2004). “Fast algorithm for detecting community structure in networks.” Available at: http://arxiv.org/pdf/cond-mat/0309508.pdf

Newing, A., G.P. Clarke and M. Clarke. (2015). “Developing and Applying a Disaggregated Retail Location Model with Extended Retail Demand Estimations.” Geographical Analysis, 47, 219-239.

National Planning Policy Framework (NPPF). (2012). Available at: http://planningguidance.communities.gov.uk/blog/policy/achieving-sustainable-development/annex-2-glossary/

OpenData. (2015). Accessed: 05/09/2016, Available at: http://opendata-aha.net/retail-places-new-release-for-2015/

Pons, P. and M. Latapy. (2005). “Computing communities in large networks using random walks.” Available at: http://arxiv.org/pdf/physics/0512106v1.pdf

Rogstad, L. and M. Dysterud. (1996). “Land use statistics for urban agglomerations: development of a method based on administrative records and the use of geographical information systems.” Statistical Journal of the United Nations Economic Commission for Europe, 13(4), 413-417.

Scharl, T. and F. Leisch. (2006). “The stochastic QT-clust algorithm: evaluation of stability and variance on time-course microarray data.” Available at: http://www.statistik.lmu.de/~leisch/papers/Scharl+Leisch-2006.pdf

Singleton, A.D., L. Dolega, D. Riddlesden and P. Longley. (2016). “Measuring the spatial vulnerability of retail centres to online consumption through a framework of e-resilience”. Geoforum, 69, 5-18.

Teller, C. and T.. Reutterer. (2008). “The evolving concept of retail attractiveness: What makes retail agglomerations attractive when customers shop at them?” Journal of Retailing and Consumer Services, 15, 127–143.

Thurstain-Goodwin, M. and D. Unwin. (2000). “Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations.” Trans. GIS 4 (4), 305-317.

Wrigley, N. and D. Lambiri. (2015). “British high streets: from crisis to recovery? A comprehensive review of the evidence.” Available at: http://thegreatbritishhighstreet.co.uk/pdf/GBHS-British-High-Streets-Crisis-to-Recovery.pdf

Xu, X., M. Ester, H.P. Kriegel and J. Sander, J. (1998). “A distribution-based clustering algorithm for mining in large spatial databases.” In Data Engineering, Proceedings., 14th International Conference pp. 324-331.

Table 1. Results from the qualitative comparison of the clustering methods in eight locations across Great Britain

Case study area	Retail center type	Preferred method
Abertillery, Wales	Small town center	KDE, Random Walk
Bristol, England	Large urban area	DBSCAN
Cardiff, Wales	City center	DBSCAN
Clapham Street, England	Large high street	DBSCAN
Glasgow, Scotland	Large city center	DBSCAN
Inverurie, Scotland	Small high street	DBSCAN
Winchester, England	Historic town center	DBSCAN
Wolverhampton, England	Regional town center	DBSCAN, Random Walk

Table 2. Optimal epsilon values used by DBSCAN in the selected study areas.

Study Area	DBSCAN epsilon (meters)
Abertillery	84
Bristol	119
Cardiff	120
Clapham Junction	70
Glasgow	70
Inverurie	120
Winchester	80
Wolverhampton	91

Table 3. Summary values of five clustering models with different standard deviation thresholds.

Models	Number of Clusters	Distribution of epsilon values (meters)					
Standard Deviation Threshold	Count	Minimum	25%	50%	Mean	75%	Maximum
0.6	2928	80	80	80	100.3	113.0	170.0
0.8	2922	80	80	80	100.3	113.0	170.0
1.0	2920	80	80	80	100.3	113.0	170.0
1.2	2923	80	80	80	100.1	113.0	170.0
1.4	2921	80	80	80	100.1	113.0	170.0

Table 4. Summary values describing the spatial distribution of the clustered locations within the Geolytix boundaries.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.68	63.97	89.81	73.99	95.99	100.00